PERGAMON

Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®

Journal of the Franklin Institute 341 (2004) 77–88

Journal
of The
Franklin Institute

www.elsevier.com/locate/jfranklin

A strategy for identifying putative causes of gene expression variation in human cancers

Sampsa Hautaniemi[a,*], Markus Ringnér[b], Päivikki Kauraniemi[c], Reija Autio[a], Henrik Edgren[d,e], Olli Yli-Harja[a], Jaakko Astola[a], Anne Kallioniemi[c], Olli-Pekka Kallioniemi[e]

[a] *Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere 33101, Finland*
[b] *Department of Theoretical Physics, Lund University, Lund, Sweden*
[c] *Laboratory of Cancer Genetics, Institute of Medical Technology, University of Tampere and Tampere University Hospital, Tampere FIN-33520, Finland*
[d] *Biomedicum Biochip Center, P.O. Box 63, University of Helsinki, Helsinki 00014, Finland*
[e] *Medical Biotechnology Group, VTT Technical Research Centre of Finland, P.O. Box 106, Turku FIN-20521, Finland*

Abstract

The majority of microarray studies focus on analysis of gene expression differences between various specimens or conditions. However, the causes of this variability from one cancer to another, from one sample to another and from one gene to another often remain unknown. In this study, we present a systematic procedure for finding genes whose expression levels are altered due to an intrinsic or extrinsic explanatory phenomenon. The procedure consists of three stages: preprocessing, data integration and statistical analysis. We tested and verified the utility of this approach in a case study, where expression and copy number levels of 13,824 genes were determined in 14 breast cancer cell lines. The procedure resulted in identification of 92 genes whose expression levels could be explained by the variability of gene copy number. This set includes several genes that are known to be both overexpressed and amplified in breast cancer. Thus, these genes may represent an important set of primary, genetically altered genes that drive cancer progression.

Keywords: Bioinformatics; Cancer; Data analysis; Statistics

*Corresponding author. Tel.: +358-3-3115-3878; fax: +358-3-3115-3817.
E-mail address: sampsa.hautaniemi@tut.fi (S. Hautaniemi).

0016-0032/$30.00
doi:10.1016/j.jfranklin.2003.12.005

1. Introduction

DNA microarray technology has become a standard tool in molecular biology by enabling measurement of expression levels of thousands of genes simultaneously. Microarrays are found to be useful in various areas of cancer research, for example, in finding significant biological differences between samples from different kinds of tumors [1,2]. Most of the published studies describe genes and clusters of genes that are able to discriminate between two or more different tumor types, or between two or more biological treatments. However, in a typical study utilizing gene expression data, causes for gene expression alterations remain unknown since only gene expression data are utilized. Thus, very little information is available on the underlying causes of the variability seen in gene expression patterns.

We are interested in attributing the variability of expression levels of genes across multiple samples to either intrinsic (DNA sequence, biological role) or extrinsic features (measured with another method) of the genes. In this study[1] we present a general and systematic procedure, which can be used in explaining gene expression variation across a set of experiments or samples.

The procedure consists of three stages: preprocessing, data integration and statistical analysis. Each stage can be modified to accommodate the specific purpose of the experiment. The heart of the procedure is data integration, where data from an explanatory phenomenon are combined with the gene expression data. In the statistical analysis stage the genes are ranked so that the top of the resulting list contains genes whose expression levels are very likely caused by the explanatory phenomenon.

We assume that for each gene expression value there is a corresponding explanatory value. The explanatory value could be another microarray measurement, gene ontology term, promoter sequence, etc. The procedure allows missing values, so actually we assume that for each gene expression value, there is the possibility to obtain an explanatory value.

In this study we first briefly review the microarray technology, which is followed by a description of the procedure for identifying genes whose expressions are altered due to the explanatory phenomenon. In the case study section we show an example where we have utilized copy number levels as explanatory data for gene expressions. Copy number alterations are considered to be one of the most influential factors for altered gene expression levels in cancer.

2. Microarray technology

Microarrays are typically used for measuring relative gene expression levels. In general, a microarray experiment proceeds as follows. Total RNA or mRNA is

[1] A preliminary version of this paper appeared in proceedings of Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, NC, USA, October 12–13, 2002.

extracted from test and reference samples. The samples are labeled with different fluorescent dyes, for example, test sample with Cy3 and reference sample with Cy5. The two samples are combined and hybridized to a microarray slide onto which probes have been immobilized in defined array format. The fluorescence intensities are then measured providing a pseudocolored image where red spots indicate upregulation of the corresponding genes in the test sample, while green spots indicate downregulation in the test sample. Typically, the ratio of the average red and the average green intensities within a spot are used as a value for the expression level of the corresponding gene. The schematic of a microarray experiment is illustrated in Fig. 1.

In the case study section the explanatory data are gene copy number levels. In order to measure copy number levels we have utilized the comparative genomic hybridization (CGH) technology [3–5]. Microarray slides for a CGH experiment are fabricated in a similar fashion like cDNA microarray slides. However, in the CGH microarray experiment, labelled DNA is hybridized on the microarray slide rather than labelled and reverse transcribed RNA. By this means CGH microarray experiments enable measurements of copy numbers of thousand of genes throughout the genome in parallel.

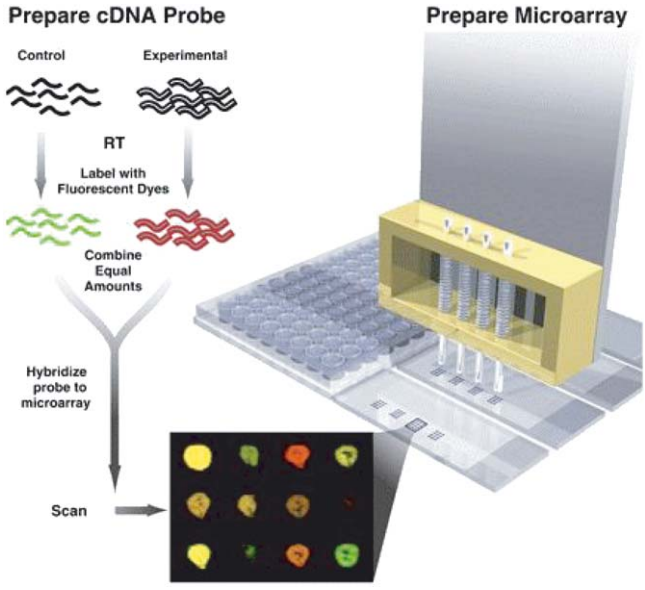


Fig. 1. Schematic of cDNA microarray experiment. A microarray slide is fabricated by printing cDNA fragments to the microarray. Test and reference samples are labeled with different fluorescent dyes and hybridized to the microarray. Hybridization is followed by washing and scanning. From the resulting image relative gene expression levels can be quantified.

3. Systematic procedure for explaining gene expressions

In order to identify the impact of an explanatory variable on gene expression we utilize three stages: preprocessing, integration and statistical analysis. The result of the procedure is a list where genes are ranked according to the significance the explanatory variable has on the gene expression level. A schematic of the steps is illustrated in Fig. 2.

3.1. Preprocessing

Preprocessing includes both within-slide and between-slide normalizations. As preprocessing is strongly dependent on the purpose of the experiment and the way the experiment has been conducted, there are no general guidelines for choosing appropriate preprocessing methods. We do not make assumptions regarding the applied preprocessing methods, so any sensible preprocessing method is applicable.

The input of the preprocessing stage is both a gene expression matrix ($\mathbf{R} \in \mathbb{R}^{n \times N}$) and an explanatory data matrix ($\mathbf{E} \in \mathbb{R}^{n \times N}$), where n is the number of genes included in the study and N is the number of samples. Moreover, the rows and columns are ordered identically in **R** and **E**, i.e. gene in ith row in **R** is the same gene as in ith row in **E**. Output of the preprocessing stage is preprocessed **R** and **E**.

3.2. Data integration

The core of the procedure is the data-integration stage, in which explanatory data and expression data are integrated. In essence, data integration is done in two phases. The purpose of the first phase is to quantize explanatory data into predetermined classes. This phase is referred to as *labeling*. In the second phase, gene expression data and quantized explanatory values are used in order to compute a value that describes how well the explanatory value can explain gene expression. This

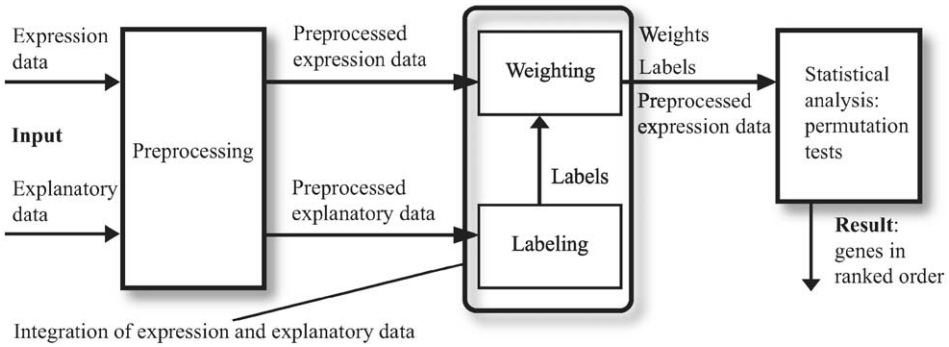


Fig. 2. Schematic of the procedure. Gene expression and explanatory data are first preprocessed and then integrated. Finally, statistical significance is computed using weights, labels and gene expression data.

phase is referred to as *weighting*. The output of the data-integration stage is a weight value for each of the genes included in the study.

3.2.1. Labeling

In the labeling phase, the explanatory data are divided into a predefined number of groups. In this study, we assume that explanatory data are divided into two groups. The first group (G_0) consists of values in **E** that are not attributable for explaining gene expression variation. The second group (G_1) consists of values in **E** that are attributable for explaining gene expression variation. For example, in our case study we are studying the impact of the increased gene copy number on gene over-expression. Values in **E** that represent increased copy numbers are grouped to G_1 and all the other values are grouped to G_0.

The result of the labeling phase is an index matrix $\mathbf{I} \in \mathbb{Z}^{n \times N}$, where entries are zeros and ones corresponding to the groups G_0 and G_1, respectively. **I** may contain missing values if there are such in the explanatory data set.

The labeling phase can be executed in several ways. For example, an intuitive approach to do labeling is to utilize a statistical test, in which case hypothesis H_0 is "this observation does not belong to the group that explains gene expression variation" and H_1 is the complement of H_0. Another approach is to utilize clustering, for example, k-means clustering or to utilize a priori knowledge, for example, that approximately 5% of the values belong to G_1.

3.2.2. Weighting

In the weighting phase, **E** and **I** are used for getting a value (W_i) that describes how well the explanatory value can explain the gene expression for the ith gene ($i = 1, \ldots, n$). As in this study we have divided the explanatory data into two groups, **I** is used to divide the values in **E** into two groups. Then for each gene, W is computed by measuring how far expression values, whose index is zero are from those expression values whose index is one. Again, there are almost arbitrarily many ways to measure how far gene expression values whose explanatory values belong to G_1 are from those values whose explanatory values belong to G_0. However, the algorithm should result in large weights when separation between the groups is good. One simple weighting method is to compute the mean of the expression for both groups and calculate their difference.

3.3. Statistical analysis

A large W does not necessarily mean that the gene's expression variation can be explained by the explanatory phenomenon, since, depending on the algorithm chosen in the labeling and weighting phases, some misclassifications are likely to occur. Therefore, the final stage in our procedure is to compute statistical significance for the weighting.

In this study, we used permutation tests to test if a large weight for a gene is really due to the explanatory phenomenon. For a profound discussion on permutation tests in general and in biomedical research we refer to References [6,7], respectively.

Table 1
Pseudo-code for computing α-value for one gene

```
IN: ith row of I (I_i) and of R (R_i), number of permutations n_p.
OUT: α-value for ith gene.
counter := 0
W_i = ComputeWeight(I_i, R_i)
Repeat n_p times
        PermLabels = Permute(I_i)
        W_new = ComputeWeight(PermLabels, R_i)
        if |W_new| > |W_i|
                counter := counter + 1
        end
end
α = counter/n_p
```

Here the permutation test is executed by permuting the ith gene's entries in **I** and computing a new W_i. Permutation results in random groups whose sizes are the same as in the original grouping. The permuted labels are used for computing a new weight, which is compared to the original weight computed in the data-integration stage. The result of the statistical test is an α-value that denotes the probability that H_0: "large weight is due to random event" is erroneously rejected. Pseudo-code for assessing the α-value for ith gene is illustrated in Table 1.

Table 1 contains two function calls. The first function ComputeWeight computes the weight for a given gene using an index vector and corresponding gene expression values. The second function Permute results in a randomly permuted index vector.

After α-values are computed for all genes included in the study, the genes are ranked according to their α-values in increasing order.

4. Case study

Most functional genomic studies of cancer and other diseases are based on assessing steady-state expression levels of thousands of genes by cDNA microarrays. Our aim is to identify underlying causes of these patterns, a process that would eventually enable a mechanistic understanding of the dysregulation of gene expression in cancer. One important determinant of gene expression in cancer is variation in gene copy number (by e.g. gene amplification). Copy number changes are common aberrations in cancer and are known to involve genes that play a crucial role in the development and progression of cancers [4,8]. In addition to offering a basis for understanding the pathogenesis of cancer, copy number alterations are considered to be one of the most influential factors for altered gene expression levels. Copy number levels can be measured by CGH microarrays [5].

We have used cDNA printed microarrays containing 13,824 genes to determine both gene expression and copy number levels in 14 breast cancer cell lines. The breast cancer cell lines included to this study were BT-20, BT-474, HCC1428, Hs578t,

MCF7, MDA-361, MDA-436, MDA-453, MDA-468, SKBR-3, T-47D, UACC812, ZR-75-1, and ZR-75-30. Thus, $\mathbf{R}, \mathbf{E} \in \mathbb{R}^{13824\times 14}$. Materials and methods for the CGH and cDNA experiments are given in [9]. After hybridization and scanning, both cDNA and CGH ratios were computed using mean intensities.

Biological aspects of this case study are elaborated in [9], where labeling was done so that 5% of all CGH-values were considered to be amplified and thereby belonging to G_1, while all the other values belonged to G_0. Here we have utilized a systematic way for labeling and rerun the analysis.

4.1. Preprocessing for case study

The overall quality of the data included in to this study was good (data shown in [9]) and therefore we performed within-slide normalization for both cDNA and CGH experiments using linear calibration method as introduced in [10].

After calibration we performed quality filtering of microarray spots as follows. In the cDNA data we discarded all spots whose mean red (test sample) and green (reference sample) intensities were under 100 fluorescent units. Moreover, we discarded spots with area smaller than 50 pixels. In the CGH data we discarded all spots whose green intensity was below 100 fluorescent units. We want to point out that we did not exclude any gene from further analysis; data from spots of poor quality were discarded and the corresponding ratios treated as missing values in the subsequent analysis. The CGH and cDNA calibrated intensity ratios were log-transformed and normalized using median centering of the values in each cell line. Furthermore, cDNA ratios for each gene across all 14 cell lines were median centered.

4.2. A systematic approach for labeling used in the case study

As described in Section 3.2, labeling is one of the key-factors affecting the outcome of the procedure. Thus, labeling should be reliable and systematic. However, changes in copy number levels are typically defined by using arbitrary cut-points. Previously, we have developed a freely available MATLAB toolbox, *CGH-Plotter*, for systematic identification of copy number changes in microarray data [11]. Here we have integrated the CGH-Plotter to the procedure explained in Section 3.2 and in this section we will briefly review the CGH-Plotter.

Copy number changes such as amplifications and deletions usually span a sizable region of the genome ranging from hundred kilobases to few megabases. Therefore, copy number ratios derived from a particular sample, when sorted in chromosomal order along the genome, can be considered as a signal that contains regions of constant levels which are to be identified. Chromosomal regions with constant levels above the baseline are called amplicons, while constant levels below the baseline are called deletions. The main purposes of the CGH-Plotter are to allow the user to plot CGH copy number data as a function of the position of the genes along the human genome and to rapidly determine the exact locations of amplicons and deletions.

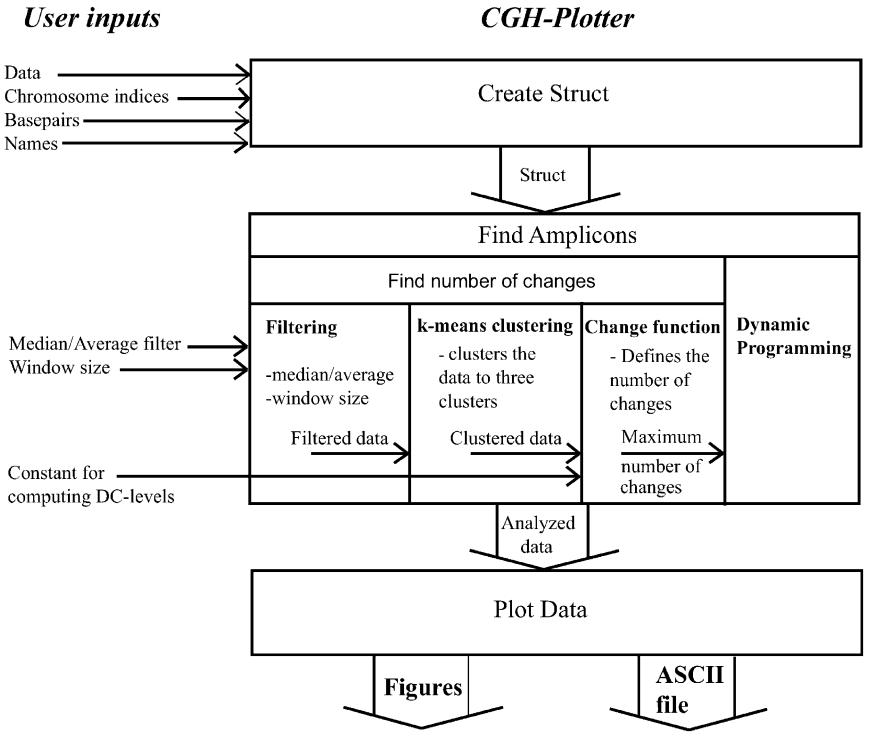


Fig. 3. Overall view and data flow for the CGH-Plotter. The user inputs the CGH-data, chromosome indices, cumulative base-pairs and names of the samples. Parameters needed are type of the filter and size of the window, and constant that affects the total number of amplicons and deletions. Here, the user needs also to determine the threshold for amplicon region to mark values in G_1.

The CGH-Plotter analyzes the data in three stages, which are also illustrated in Fig. 3:

(1) *Filtering*: In order to reduce the effect of noise, the CGH-Plotter first filters the data. The size of the filter window and the type of the filter are inputted in the graphical user interface.
(2) *k-means clustering*: The CGH-Plotter clusters the filtered data with the k-means clustering algorithm to find the maximum number of amplicons and deletions in each chromosome. The number of clusters (k) used is three denoting "amplified", "deleted" and "baseline" regions.
(3) *Dynamic programming*: A dynamic programming function determines the constant levels in the CGH-data. Here it is assumed that the CGH-data can be approximated by a constant together with additive Gaussian noise. Thus, the CGH-data can be understood as a signal having constant levels, which are identified with a dynamic programming algorithm. It is assumed that the number of constant levels is known. The dynamic programming method uses the

Markov property and identifies change points of the constant levels by minimizing the mean square errors for all combinations of CGH-ratios.

The CGH-Plotter returns the heights of the amplicons, which are used to determine whether a particular gene is considered to be amplified as follows. All genes belonging to the amplified region determined by the CGH-Plotter with height more than 1.4 were grouped to G_1. Thus, for our case study data set, genes with CGH ratio > 1.75 (representing 1.4% of all CGH-values across the experiments) were considered to be amplified. An example of the CGH-Plotter figure and output file is given in Fig. 4.

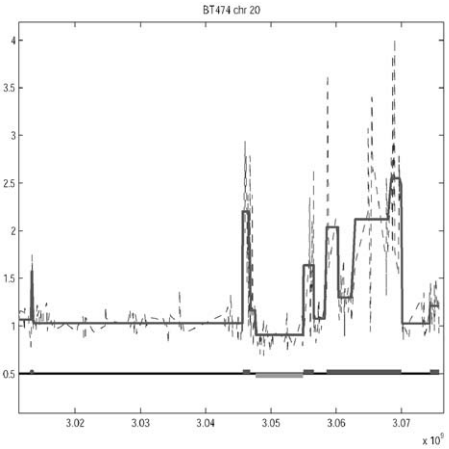
4.3. Weighting used in the case study

In order to compute W for ith gene, we utilized signal-to-noise statistics [12]:

$$W_i = \frac{m_1 - m_0}{(\sigma_0 + \sigma_1)}, \tag{1}$$

where m_1, σ_1 and m_0, σ_0 denote the sample means and sample standard deviations for the expression levels for amplified and nonamplified samples, respectively.

Signal-to-noise statistics results in a large weight if the means of the groups are far away from each other and standard deviations within the groups are small.



Microsoft Excel - Boundaries_chr20

File Edit View Insert Format Tools Data Window Help Acrobat

	A	B	C	D	E	F
1	CHROMOSOME20	SKBR3	MCF7	BT474	BT20	MDA361
2						
3	1 Type	Amplicon	Amplicon	Amplicon	Deletion	Deletion
4	1 Number of clones	9	10	9	8	16
5	1 Start	10711	10710	10711	10695	10695
6	1 End	10719	10719	10719	10702	10710
7	1 Start Basepair	3013370320	3013218324	3013370320	3011240006	3011240006
8	1 End Basepair	3013370320	3013370320	3013370320	3011424806	3013218324
9	1 Height	1.7251	1.3247	1.5765	0.74867	0.9361
10	1 Max/Min	2.1341	1.4997	1.7483	0.57769	0.76154
11						
12	2 Type	Amplicon	Amplicon	Amplicon	Amplicon	Amplicon
13	2 Number of clones	49	13	11	8	9
14	2 Start	10838	10890	10838	10711	10711
15	2 End	10886	10902	10848	10718	10719
16	2 Start Basepair	3045696934	3052383073	3045696934	3013370320	3013370320
17	2 End Basepair	3050280636	3053426980	3046560263	3013370320	3013370320
18	2 Height	1.3474	1.2754	2.1995	1.738	1.3475
19	2 Max/Min	1.8628	1.8892	2.9362	1.9891	1.5841

Fig. 4. On the left, copy number ratios and the chromosome boundaries resulting from the analysis of chromosome 20 of breast cancer cell line BT474. Data are plotted against cumulative base-pairs, which tell the actual location of each gene in the genome. y-Axis depicts the copy number ratios between test sample and reference. Original CGH-data are plotted with thin dotted line and amplicon boundaries with a thick continuous line. Bars below the data correspond to the amplicons and deletion. On the right, partial results from the analysis of five breast cancer cell lines in ASCII-file. Entries in the result file are: type of the copy number change, number of clones in the change region, start and end points of the amplicon or deletion, height of the amplicon or deletion and maximum and minimum ratio inside the amplicon or deletion.

4.4. Results

In order to obtain statistical significance for the weights we utilized permutation test as described in Table 1. For each gene we performed 10,000 permutations and obtained an α-value. A low α-value indicates a strong association between gene expression and gene amplification.

The procedure resulted in identification of 92 genes whose α-value was below 0.05. This list included the *HOXB2* and *HOXB7* genes, which were validated with reverse transcription polymerase chain reaction (RT-PCR) and fluorescence in situ hybridization (FISH), as well as the *EGFR* and *ERBB2* genes, which are known to be over-expressed and amplified in breast cancer [9]. The list included also a few genes that were not included in the original list given in [9]. Validation of the role of these genes in breast cancer is in progress, and beyond the scope of this study.

We used Gene Ontology annotations [13] to investigate what biological processes the genes on the list of 92 genes take part in. The most frequent annotations include processes such as cell growth, nucleotide and protein metabolism, cell communication and regulation of transcription.

We applied the procedure to also identify genes that are both deleted and underexpressed. Labeling was done with CGH-Plotter so that deleted areas whose height was 0.8 or below were considered as deletions. The majority of the genes found were deleted in only one cell line, and not a single gene was deleted in four or more cell lines. Therefore, the results from that simulation may not be reliable and require validation beyond the scope of this study.

5. Discussion

We have shown a systematic approach for identifying genes whose expression levels are significantly influenced by an explanatory phenomenon. Since genes that undergo amplification or other "genetic change" in cancer may be the primary "driver genes" of cancer development and progression, the procedure enabled us to quickly identify a small subset of genes for further analysis. This approach is therefore highly valuable in trying to prioritize and simplify the most essential gene expression information in cancer.

The crucial phase in our strategy is the labeling. If the labels in $\mathbf{I}$ are erroneous, they cannot be compensated in α-value computation. However, permutation tests could be used in assessing statistical significance to labels in a similar fashion with the procedure described in Section 3.3: For each label vector choose randomly n_p genes and compute new weights. If new weight is larger than the original weight, increase counter by one. In the end, the counter is divided by n_p. The resulting value, β-value, tells the probability that H_0: "label vector induces a unique weight" is erroneously rejected. Thus, β-value reflects the statistical significance of the labeling.

In our case study, we applied the method for a breast cancer data set and identified 92 genes whose expression levels potentially were due to an underlying gene amplification event in cancer. This amount is lower than in the original study [9] due

to fact that the threshold used for considering a CGH ratio to be amplified was higher in this study. However, the procedure presented here benefits from the systematic detection of the amplified areas, which may lead to more reliable results than in the original study. The majority of the genes identified here were the same as in [9] including genes that were known to be both over-expressed and amplified in breast cancer.

In the analysis of reduced expression levels it is very hard to distinguish whether the expression ratio is small due to biological reasons or noise. Signal-to-noise statistics consist of means and standard deviations, which are known to be sensitive to noise. Therefore, in the analysis of low-level expression changes, more has to be assumed about the data and such analysis would constitute an interesting topic for further study. For example, machine learning algorithms such as support vector machines or learning vector quantization may turn out to be useful when detecting genes that are both deleted and underexpressed.

The strategy presented here is also applicable to studies aiming to identify differentially expressed genes among two conditions. In this case, **I** is formed using two conditions, such as tumor samples (G_1) and healthy samples (G_0), rather than using explanatory data set. Thus, W_i is large if gene expression levels of the ith gene in G_1 are far away from expression levels in G_0. That is, the ith gene is differentially expressed among two conditions. Finally, statistical significance for each weight can be computed using permutation tests as presented in this study.

In summary, we have developed a procedure that could be used in studies where the underlying causes of gene expression variations are examined. When we applied the procedure to explain overexpression in a breast cancer study, the procedure was able to identify high-impact primary candidate gene targets for development of therapies and for sub-classification of breast cancer.

Acknowledgements

This work was supported in part by the Academy of Finland, the Emil Aaltonen Foundation and the Swedish Research Council.

References

[1] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J.H. L. Lu Jr.,, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (6769) (2000) 503–511.

[2] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, T.R. Golub, Multiclass cancer diagnosis using tumor gene expression signatures, Proc. Nat. Acad. Sci. U.S.A. 98 (26) (2001) 15149–15154.

[3] A. Kallioniemi, O.-P. Kallioniemi, J. Piper, M. Tanner, T. Stokke, L. Chen, H.S. Smith, D. Pinkel, J.W. Gray, F.M. Waldman, Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization, Proc. Nat. Acad. Sci. U.S.A. 91 (6) (1994) 2156–2160.
[4] A. Kallioniemi, O.-P. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman, D. Pinkel, Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors, Science 258 (5083) (1992) 818–821.
[5] J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, P.O. Brown, Genome-wide analysis of DNA copy-number changes using cDNA microarrays, Nature Genet. 23 (1) (1999) 41–46.
[6] P. Good, Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, Springer Series in Statistics, 2nd Edition, Springer, Berlin, 2000.
[7] J. Ludbrook, H. Dudley, Why permutation tests are superior to t and F tests in biomedical research, The Amer. Statist. 52 (2) (1998) 127–132.
[8] O. Monni, M. Bärlund, S. Mousses, J. Kononen, G. Sauter, M. Heiskanen, P. Paavola, K. Avela, Y. Chen, M.L. Bittner, A. Kallioniemi, Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer, Proc. Nat. Acad. Sci. U.S.A. 98 (10) (2001) 5711–5716.
[9] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringnér, G. Sauter, O. Monni, A. Elkahloun, O.-P. Kallioniemi, A. Kallioniemi, Impact of DNA amplification on gene expression patterns in breast cancer, Cancer Res. 62 (21) (2002) 6240–6245.
[10] Y. Chen, E. Dougherty, M. Bittner, Ratio-based decisions and the quantitative analysis of cDNA microarray images, J. Biomed. Opt. 2 (4) (1997) 364–374.
[11] R. Autio, S. Hautaniemi, P. Kauraniemi, O. Yli-Harja, J. Astola, M. Wolf, A. Kallioniemi, CGH-Plotter: MATLAB toolbox for CGH-data analysis, Bioinformatics 19 (13) (2003) 1714–1715.
[12] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537.
[13] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology. Nature Genet. 25(1) (2000) 25–29.